Title: PRECOG Recruitment Task, IIIT Hyderabad | NLP Theme
Report Author: Satya Swaroop Gudipudi (2022900024)

Representations for Words, Phrases, Sentences

Preprocessing:

Following text preprocessing is performed as cleaning step over the corpus:

- replace_url to replace urls with URL
- replace_hashtags to replace with HASHTAG
- replace_email to replace with EMAIL
- replace_mentions to replace with MENTION
- replace_numbers to replace with NUMBER
- remove_abbrevations to replace possessive words with their extended representations
- remove_punctuation to remove the punctuations and it can be replaced with PUNCT

Word Similarity:

Constrained

WordNet corpus is used as constrained datasource as it is a large lexical database of English. WordNet groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets.

Approach:

From the test data SimLex gathered the word pairs and extracted the synsets, hypernyms, hyponyms of these words from wordnet corpus.

corpus size: 327419 sentences
vocab size: 27952
Created co-occurrence matrix with window size 2 and applied TruncatedSVD to get dense representations for words. The cosine similarity between word pairs is scaled to meet the SimLex scale.
Following are the results obtained:

S.no	Description	MSE	Spearman
1	Word representation with co-occurrence window size 1	7.52	0.06
2	Word representation with co-occurrence window size 2	7.59	0.075

Observation:
As the dataset is constrained, training neural networks would not be able to generalize.
The constructed co-occurrence matrix is sparse, more diverse dataset could result in better representation.
As per above results with spearman correlation less it indicates the underlying word representations are not capturing the semantic information effectively.

Unconstrained:

Since as per task allowed to use any data or model. Used ELMo pretrained model and ELMo is good at generating semantic rich dense representation for words.

Experiment Description	RMSE	Spearman
ELMo pretrained model	3.97	0.41

Observations:

Error analysis shows that for antonyms the squared error is high.
From the results, it can be seen that the embeddings capture some aspect of semantic similarity and there is moderate correlation as per Spearman coefficient. This suggests ELMo is good at generating word embeddings when context given.
A more appropriate embeddings could be static embeddings which do not change with surrounding context like Word2vec, Fasttext and Glove embeddings.

Phrase Similarity:

Approach:
The word2vec embeddings for each word in phrase are computed and averaged to form a fixed size representation for the phrase of any length. On top of these averaged embeddings various supervised models are applied. Following are the results:

S.no	Model Description	Accuracy
1	SVM	0.363
2	Neural Network	0.42

Observations:

Since the dataset is balanced Accuracy metric is choose.
The average of embeddings could be resulting in information loss and not helping the classification models in better classification.

Sentence Similarity:

Approach:
The Word2vec embeddings of the sentences are fed as input to the Siamese BiLSTM network. That contains two identical sub-networks which takes in the two input sentences separately and process them through BiLSTM layer.
Architecture:
1. Shared Weights: The same bilstm layer with the same parameters (weights) is used to process both sentences (sentence1 and sentence2), which is a hallmark of Siamese architectures.
2. Symmetric processing: The forward_once method is used to process both inputs in a symmetric manner.

S.no	Model Description	Accuracy	ROC-AUC
1	Siamese BiLSTM	0.536	0.502
2	SBert Finetuning	0.90	0.96

Observations:
The transformer based SBert(Simaese based) Finetuning seems to better capture the semantic information between the sentences better than the BiLSTM network.

*ChatGPT is used as code assistant for optimizing the code.